J Mol Model (2013) 19:1355–1367
DOI 10.1007/s00894-012-1695-6

ORIGINAL PAPER

A solvated ligand rotamer approach and its application in computational protein design

Xiaoqiang Huang · Ji Yang · Yushan Zhu

Received: 12 September 2012 / Accepted: 13 November 2012 / Published online: 29 November 2012

Abstract The structure-based design of protein–ligand interfaces with respect to different small molecules is of great significance in the discovery of functional proteins. By statistical analysis of a set of protein–ligand complex structures, it was determined that water-mediated hydrogen bonding at the protein–ligand interface plays a crucial role in governing the binding between the protein and the ligand. Based on the novel statistic results, a solvated ligand rotamer approach was developed to explicitly describe the key water molecules at the protein–ligand interface and a water-mediated hydrogen bonding model was applied in the computational protein design context to complement the continuum solvent model. The solvated ligand rotamer approach produces only one additional solvated rotamer for each rotamer in the ligand rotamer library and does not change the number of side-chain rotamers at each protein design site. This has greatly reduced the total combinatorial number in sequence selection for protein design, and the accuracy of the model was confirmed by two tests. For the water placement test, 61 % of the crystal water molecules were predicted correctly in five protein-ligand complex structures. For the sequence recapitulation test, 44.7 % of the amino acid identities were recovered using the solvated ligand rotamer approach and the water-mediated hydrogen bonding model, while only 30.4 % were recovered when the explicitly bound waters were removed. These results indicated that the developed solvated ligand rotamer approach is promising for functional protein design targeting novel protein–ligand interactions.

Keywords Protein–ligand interaction · Solvated ligand rotamer · Water-mediated hydrogen bonding · Protein–ligand interface · Computational protein design

X. Huang · J. Yang · Y. Zhu (✉)
Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
e-mail: yszhu@tsinghua.edu.cn

Introduction

Protein–ligand interfaces always contain some specifically bound water molecules, bridging ligands and protein side chains via hydrogen bonds. Water molecules can play important roles in the active regions of proteins, being useful for binding between protein and ligand [1–4], and even functioning in catalysis [5]. Luque and Freire [6] found that considering buried native water molecules as part of the ligands could improve the prediction of the binding enthalpy of small molecules. In protein–protein interface design, Kortemme et al. [7] found that the experimentally determined residue conformations differ from computational conformations if water-mediated hydrogen bonding in the interface is not considered. However, water-mediated hydrogen bonding networks cannot be described accurately by continuum solvent models, although these are widely used in the field of computational protein design and, as a consequence, the energy contributions of these networks are neglected. Jaramillo and Wodak [8] compared several implicit solvation models for computational protein design and concluded that the computational protein design problem is still a great challenge for implicit solvation models. The energetics of water-mediated hydrogen bonding networks could be illustrated more accurately by discrete solvent molecules rather than the continuum solvent. A straightforward strategy to overcome this problem is to incorporate a large number of solvent molecules explicitly into the computational protein design protocol, similar to the method used for molecular dynamic simulation [9, 10], but that is a formidable computational challenge and is still intractable for computational protein design. As computational protein design has been developed to construct protein variants with novel functions, such as artificial enzymes designed for unnatural reactions [11–13], it is of great significance to investigate the specific water molecules at protein–ligand

interfaces and the corresponding modeling approaches within the computational protein design framework.

As bound water at the interface affects the binding between ligand and protein, some modeling approaches for water-mediated hydrogen bonding networks have been developed in the popular docking algorithms for high throughput virtual screening in the drug discovery field. GOLD [14] has been modified to consider water molecules by allowing each water molecule to switch on and off within the target structure. This approach has shown some success, but can deal only with cases where the positions of water molecules are known in the crystal structure, and thus water molecules lost in the crystal structure cannot be estimated. AutoDock [15] deals with water molecules by considering multiple targets within one calculation, some with water and some without. This method once again makes use of the known water molecules but fails in predicting lost water molecules. FlexX [16] treats water molecules as spheres, and an ensemble of favorably placed water molecules is predicted and docked into the binding site. Glide [17] works similarly by docking explicit water molecules into the binding site for each energetically competitive ligand pose. Two of the techniques, FlexX and Glide, take the lost water molecules into account. Rossato et al. [18] presented a directional approach (AcquaAlta) to match predicted water positions with experimental ones, and the method can be applied to protein–ligand docking where the generated water molecules could increase the probability of finding a bioactive pose and facilitate a more reliable ranking and scoring of the docked poses. Huggins and Tidor [19] developed a method to systematically place the structural water molecules around the ligand molecule instead of the protein where the explicit water molecules were found to be able to improve the scoring of protein–ligand interactions.

In the above mentioned protein–ligand docking algorithms, the identities and conformations of amino acids in the binding sites are known ahead of docking. Therefore, these approaches for modeling specific water molecules at the interface cannot be used directly in the computational protein design framework since the protein side chains never exist before the design ends. To resolve this dilemma, Jiang et al. [20] developed a solvated side-chain rotamer approach by appending optional water molecules to the end of the polar groups of amino acid rotamers. This approach can accurately predict the positions of water molecules at protein–protein interfaces. This method has been used in protein–protein interface design, and improves the prediction of amino acid identities in a series of sequence recapitulation tests. However, this approach can always greatly increase the total combinatorial number because a large number of solvated rotamers are produced based on only one normal side-chain rotamer, therefore this approach is not practical in a deterministic optimization method-based computational protein design framework.

This work presents a statistical analysis of hydrogen bonds in the active regions of 42 protein–ligand complex structures; and a large number of water-mediated hydrogen bonds were observed in highly buried regions. This observation confirms that structural water molecules are of great significance for binding between proteins and ligands. The solvated rotamer approach for specific water molecules comes from the observation that there are some centroid positions for water molecules around polar groups of amino acids [21, 22]. A novel modeling approach was developed to generate specific waters on positions spanning from the hydrogen-bonding groups of ligand in the environment of the protein–ligand interface; a large number of inappropriate waters were removed from the complete list given by the method of Huggins and Tidor [19]. A solvated ligand rotamer approach was further developed to construct an extended ligand library where each normal ligand rotamer produces only one additional solvated rotamer; therefore, the huge total combinatorial number in sequence selection can be effectively circumvented. This library and a simple free energy function to describe the water-mediated hydrogen bonding were added into our PROtein Design Algorithmic (PRODA) package [23–25], and their performance was tested on two tests: (1) the recovery of the explicit water molecules at the protein-ligand interfaces, and (2) recapitulation of active region sequence for receptor design.

Methods and materials

Datasets

Two subsets of data from the Brookhaven Protein Data Bank (PDB) [26] were used. The first set, which is composed of 42 proteins with resolution 2.0 Å or better and R-factor of 20 % or better, with 15 % sequence identity or less, was used to investigate the hydrogen bonds within the protein active site regions. The PDB identities of these 42 entries are listed in Table 1. Of the 42 entries, 25 cases were selected from the dataset used for our H-DOCK [24] testing. The remaining 17 cases were chosen from the dataset used for the statistical analysis of the buried hydrogen bonds by McDonald and Thornton [27]. Only one of the subunits in each protein structure was used when there were multiple subunits, and only one or two ligands were considered according to each specific test case when multiple ligands existed in the complexes. All experimentally determined water molecules were considered in the statistical analysis of hydrogen bonds in active regions. The second set including five PDB entries shown in Table 2 was selected from

Table 1 The 42 protein–ligand complex structures used for hydrogen bonding statistical analysis

PDB	Chain[a]	Ligand	Resolution(Å)	R-factor	PDB	Chain	Ligand	Resolution(Å)	R-factor
1ABF		FCA307	1.9	0.13	2AZA	A	SO4135	1.8	0.16
1APW	E	Chain I[b]	1.8	0.13	2CSC		MLT702	1.7	0.19
1BZM		MZM262	2.0	0.19	2GBP		BGC310	1.9	0.15
1CBX		BZS500	2.0	0.17	2OZ9	R	SO4602	1.7	0.18
1CTF		SO41	1.7	0.17	2POR		CBE545	1.8	0.19
1DRF		FOL187	2.0	0.19	2QWG		G28800	1.8	0.18
1ELA		0Z1256	2.0	0.19	2SAR		3GP98	1.8	0.18
1FKF		FK5108	1.7	0.17	2TMN		0FA332	1.6	0.18
1GOX		FMN370	2.0	0.19	2TRX		MPD607	1.7	0.17
1HSL	A	HIS239	1.9	0.20	3CHY		SO4401	1.7	0.15
1HVR	AB	XK2263	1.8	0.19	3CLA		CLM221	1.8	0.16
1LZI[c]		BHG452 FUC453	1.5	0.19	3FX2		FMN149	1.9	0.20
1PII[c]		PO4453 PO4454	2.0	0.17	3GRS		PO4480	1.5	0.19
1PPH		0ZG1	1.9	0.17	4BP2		MPD301	1.6	0.19
1RBP		RTL183	2.0	0.18	4ENL		SO4444	1.9	0.15
1RGK		2AM105	1.9	0.14	4FGF		SO4147	1.6	0.16
1RHN		SO4156	2.0	0.20	4SGA	E	Chain P[b]	1.8	0.12
1SNC		THP151	1.6	0.16	5CNA	A	MMA238	2.0	0.20
1SRE	A	HAB300	1.8	0.18	7TIM	A	PGH249	1.9	0.18
1THA	AB	T33130	2.0	0.19	8ACN		NIC755	2.0	0.16
2AK3	A	AMP226	1.9	0.19	8XIA		XLS389	1.9	0.14

[a]Where multiple chains exist, the chain used for this study is indicated

[b]The ligand is a short peptide

[c]Two ligands are selected in these PDB entries

Huggins and Tidor [19] to evaluate their solvated ligand rotamer approach.

Identification of hydrogen bonds in buried active regions

The active region of the protein is represented by the residues that lie within 7.5 Å of the ligand as determined from the complex structure. The coordinates of the hydrogen atoms of the protein were calculated by PRODA [23–25] based on standard CHARMM22 [28] topology parameters. Hydrogen atoms were added to the ligand using Discovery Studio software [29]. The criteria for hydrogen bonding formation given by Baker and Hubbard [30] were used to identify potential hydrogen bonds between residue and residue, and residue and ligand. No standard criteria are available for hydrogen bonds mediated by the water molecules. Here, a polar heavy atom (mainly O or N) is defined as forming a hydrogen bond with a water molecule if the distance between the atom P and the atom W is less than 3.5 Å without any restriction on angles, where P is the polar atom, and W is the oxygen atom of a water molecule.

Table 2 Water placement results at protein–ligand interfaces

Protein (ligand)	Resolution (Å)	No. of native water molecules (total water growable positions)	No. of predicted waters (no. of correct waters)
1DF7(MTX)	1.70	4(19)	9(4)
1JIO(DEB)	2.10	3(14)	2(2)
1KI8(BVD)	2.20	3(13)	5(1)
1NNC(ZMR)	1.80	4(22)	8(2)
1VZQ(SHY)	1.54	4(14)	7(2)

For proteins, the hydrogen bonding donors considered included Asn ND2, Arg NE, Arg NH1, Arg NH2, Gln NE2, His ND1, His NE2, Lys NZ, Ser OG, Thr OG1, Trp NE1, Tyr OH and main-chain NH, and the hydrogen bonding acceptors included Asn OD1, Asp OD1, Asp OD2, Gln OE1, Glu OD1, Glu OD2, His NE2, His ND1, Ser OG, Thr OG1, Tyr OH and main-chain CO. For the ligand, the potential hydrogen bonding donors and acceptors were identified by atom type, hybridization state, functional group and specific chemical environment. These hydrogen bonding groups were divided into four types: sp^2-hybridized nitrogen, sp^3-hybridized nitrogen, sp^2-hybridized oxygen, and sp^3-hybridized oxygen. As sulfur atoms are relatively weak hydrogen bonding acceptors, they were not considered in either the protein or the ligand. Because the OH groups in Ser and Thr are rotatable and the OH group in Tyr can be turned over, these hydroxyl groups affect the number of

hydrogen bonds formed. For the experiments described herein, the OH in Ser and Thr was rotated and sampled every 10°, while the hydrogen atom of OH in Tyr was sampled at two symmetric positions on the same aromatic ring plane. The rotatable characteristic of the NH_3 in Lys was not considered. The ligand hydroxyl groups were treated in the same way as described above.

The solvent accessible surface areas (SASA) of polar atoms at the protein–ligand interface were calculated using a numerical surface calculation (NSC) algorithm [31] with a solvent probe size of 1.4 Å. The buried ratio of a heavy polar atom in the residue or the ligand was calculated based on a reference state defined as the residue i in question plus the local backbone atoms CA($i-1$), C($i-1$), and O($i-1$) in the preceding residue and N(i+1), H(i+1), and CA(i+1) in the following residue. It should be noted that explicit water molecules were not considered while the SASA was calculated.

Protein design at the protein–ligand interface

The sequence selection calculation in the active region of the protein–ligand interface was similar to that in the protein core redesign [23], and the code was implemented in the PROtein Design Algorithmic (PRODA) package [23–25]. The coordinates of the ligand were taken from the crystal structure and remained fixed during the whole redesign process. Only polar sites that lie within 5 Å from the ligand were chosen as designated design positions. These positions were allowed all conformations from the following polar amino acid types to be sampled: Arg, Asn, Asp, Gln, Glu, Cys, His, Lys, Ser, Thr, and Tyr. The backbone atoms and the non-optimized side-chains were held fixed at their coordinates in the crystal structure. The side-chain conformations of the designed residues come from a backbone-independent rotamer library of Xiang and Honig [32], which contains 984 rotamers for all types of amino acids. The total number of the allowed rotamers at each design position is 856 and the total number of the design sites is up to 12. Therefore the final sequence selection calculation was a huge combinatorial optimization problem that was solved by dead-end elimination [33–35] and the mixed-integer linear programming combined algorithm developed in PRODA [23].

In our former work [23], a simple free energy function was used for protein core redesign comprised mainly of a van der Waals term and a solvation term for nonpolar residues. In this work, a more complicated free energy function based on a molecular mechanics energy model and an implicit solvent model was used to characterize the interactions between the polar and charged residues accurately. Specifically, the energy was a summation of the following terms: (1) the van der Waals attractive and repulsive terms [36], where the van der Waals radii and well depths were taken from the CHARMM22 [28] parameter set, except that the van der Waals radii for polar hydrogen were scaled by 0.5 and 0.95 for other atoms in the repulsive term; (2) an explicit, geometry- and hybridization-dependent hydrogen term [37]; (3) a hydrophobic term [38] to favor the nonpolar surface area burial with a parameter of 26 cal$\,$mol$^{-1}\,$Å^{-2}; (4) a generalized Born electrostatic terms for the desolvation energy of the polar atoms upon burial after design and the screened Coulomb interaction between atoms with partial charges [39–41]; (5) the side-chain entropy loss term upon formation of the folded state [42].

Solvated ligand rotamer generation

Figure 1 illustrates the main processes for generating the solvated ligand rotamers in the framework of protein design at the protein–ligand interface, and the detailed steps are presented as follows:

Step 1. A protein–ligand interface for redesign is shown in Fig. 1a, where the optimized residues were truncated to Ala. The conformation of the ligand can be taken from the crystal structures or more generally from a ligand rotamer library, which can be produced by using the targeted small molecule placement approach [43]. However, the ligand conformations in this work were all taken directly from the crystal structures to eliminate the adverse effects of small molecule flexibility on water-mediated hydrogen bonding. It should be noted that solvated ligand rotamer generation has not been tested in cases where the conformation of the ligand bound to the protein is unknown.

Step 2. Determine all water growable positions of the ligand and add water molecules at all these positions as shown in Fig. 1b. The water growable positions are the potential sites corresponding to the polar functional groups of the ligand [19]. Water molecules were added to all these positions using the specific geometrical parameters given in Table 3, where two water molecules were associated with the sp^2 hybridized O, such as the carbonyl O, along the direction of two lone pairs, while three staggered water molecules were attached to the hydroxyl groups like serine and threonine. For the hydroxyl group like that in tyrosine, two water molecules were assumed to exist on the same plane of the aromatic ring. For the nitrogen atom, water molecules were added along the direction of the hydrogen atoms and the lone pairs.

Step 3. Remove water molecules in collision with true atoms of the protein and the ligand. Specifically, water molecules were removed if one of the following conditions was satisfied: (1) the distance between the water and the non-polar heavy atom was

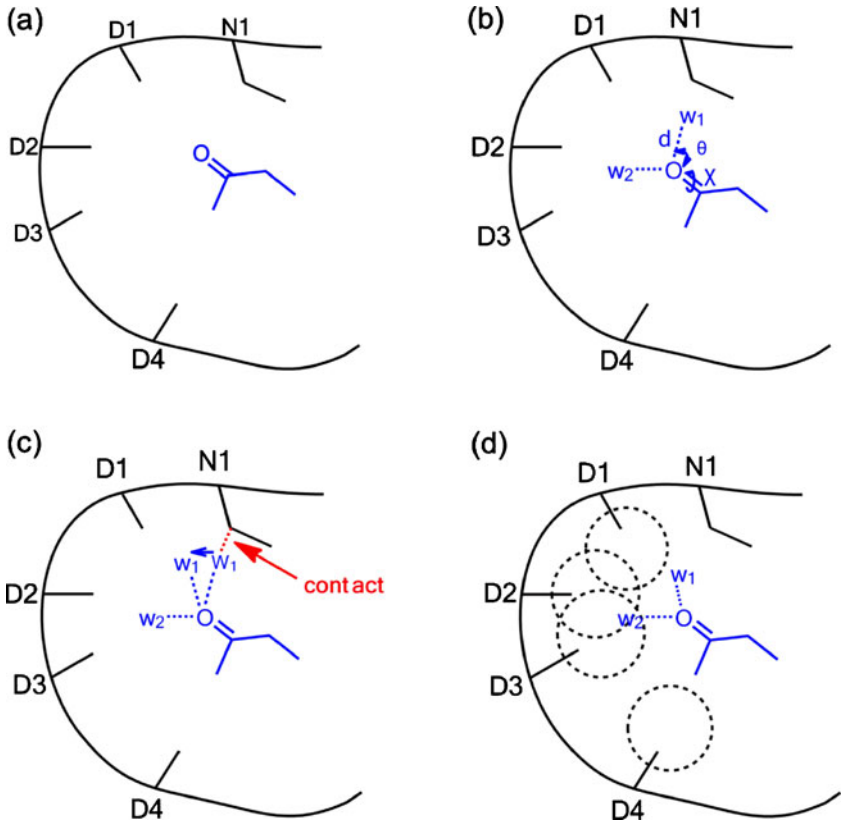
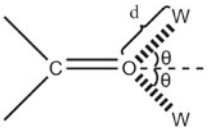
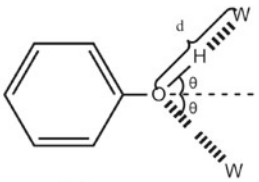
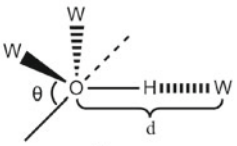
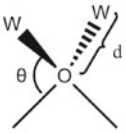
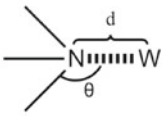
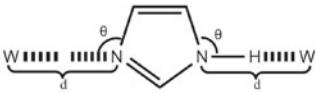


Fig. 1a–d Main generation steps in the solvated ligand rotamer approach. **a** The design site represented by D1–D4 are truncated. Other sites, such as N1, are fixed. **b** Place waters in all growable positions of the ligand. **c** Move water to a better position. **d** Water molecules are removed by clashing pseudo spheres. *Blue* Ligand, *black* protein

Table 3 Geometrical parameters for water placement based on Table I in Jiang et al. [20]

Polar group(hybridization)	Illustration of water positions	Distance (Å)	Angle θ
Carbonyl oxygen(sp2)		2.8	50°
Hydroxyl oxygen (sp2)		2.8	72.5°
Hydroxyl oxygen(sp3)		2.8	109.5°
Ester oxygen(sp3)		2.8	109.5°
Amine nitrogen(sp2/sp3)		3.0	109.5°
Aromatic nitrogen(sp2)		3.0	125.0°

<2.4 Å; (2) the distance between the water and the polar heavy atom was <1.8 Å; (3) the distance between the water and the non-polar hydrogen atom was <1.5 Å; (4) the distance between the water and the polar hydrogen atom was less than 1.0 Å; and (5) if the distance between two water molecules was <1.4 Å, the one with more clashes was removed.

Step 4. Adjust the positions of the water molecules that come into tight contact with other atoms as shown in Fig. 1c. The water molecule was moved to a new position if one of the following conditions was met: (1) the distance between the water and the non-polar heavy atom was <3.5 Å; (2) the distance between the water and the polar heavy atom was <2.6 Å; (3) the distance between the water and the non-polar hydrogen atom was <2.0 Å; (4) the distance between the water and the polar hydrogen atom was <1.5 Å; and (5) the distance between two water molecules was <2.6 Å. The principle of water moving is that the clashes will decrease and the formed water-mediated hydrogen bonding will not disappear. The specific process for water moving is implemented by identifying a new position for a water molecule within a 2.0 Å cube centered on its initial position by scanning each grid point in the cube by a step of 0.2 Å in three coordinate directions. At each step, the van der Waals repulsive energy and hydrogen bonding energy by using a set of scoring functions described in Table 4 were evaluated. The water molecules were moved to the new positions with minimum repulsive energies. The position having better hydrogen bonding energy was chosen if there were multiple positions with identical repulsive energy. A list of water molecules within contact distances were recorded and moved one by one. Once a water molecule was moved out of the contact distance, the list was updated. If no water was moved out of the contact distance of other atoms, the process was terminated.

Step 5. Remove water molecules in collision with pseudo spheres at the design sites, shown in Fig. 1d, where the pseudo spheres were attached on the CB atom along the direction from CA to CB, the distance between the CA atom and the centroid of the pseudo sphere was 2.2 Å and the uniform radius of the pseudo sphere was 3.0 Å. Water molecules were removed if the distance between the water and the pseudo sphere was less than 80 % of the sum of radii of the water and the pseudo sphere.

Step 6. Remove the water molecules in relatively empty spaces at the protein–ligand interface. The water molecules were removed if one of the following conditions held: (1) no CB atom existed within 8 Å to the water molecule; (2) no CB atom existed within 7–8 Å to the water molecule as well as the angle between CB, CA and water oxygen was <45°;

Table 4 Scoring function for water moving, where d (Å) is the distance between water oxygen and the neighboring atom, and θ (°) is the angle between water oxygen atom, polar heavy atom, and the heavy atom attached to the polar heavy atom

Scoring function	Description
$VDW = \begin{cases} 10.0 - \frac{10.0}{3.5}d & (0 \le d \le 3.5) \\ 0.0 & (d > 3.5) \end{cases}$	Used to calculate the repulsive energy between water oxygen atom and the non-polar heavy atoms
$VDW = \begin{cases} 10.0 - \frac{10.0}{2.6}d & (0 \le d \le 2.6) \\ 0.0 & (d > 2.6) \end{cases}$	Used to calculate repulsive energy between water oxygen atom and polar heavy atoms and other water oxygen atoms
$VDW = \begin{cases} 10.0 - \frac{10.0}{2.0}d & (0 \le d \le 2.0) \\ 0.0 & (d > 2.0) \end{cases}$	Used to calculate repulsive energy between water oxygen atom and non-polar hydrogen atoms
$VDW = \begin{cases} 10.0 - \frac{10.0}{1.5}d & (0 \le d \le 1.5) \\ 0.0 & (d > 1.5) \end{cases}$	Used to calculate repulsive energy between water oxygen atom and polar hydrogen atoms
$Hbond = F(d) \cdot G(\theta)$, where $F(d) = \begin{cases} 0.0 & (d < 2.6 \ \ or \ \ d > 3.5) \\ -8.5(d-2.6) & (2.6 \le d < 2.7) \\ -0.85 - 1.5(d-2.7) & (2.7 \le d < 2.8) \\ -1.00 + 0.6(d-2.8) & (2.8 \le d < 2.9) \\ -0.94 + 1.9(d-2.9) & (2.9 \le d < 3.0) \\ -0.75 + 3.0(d-3.0) & (3.0 \le d < 3.1) \\ -0.45 + 2.8(d-3.1) & (3.1 \le d < 3.2) \\ -0.17 + 1.2(d-3.2) & (3.2 \le d < 3.3) \\ -0.05 + 0.4(d-3.3) & (3.3 \le d < 3.4) \\ -0.01 + 0.1(d-3.4) & (3.4 \le d \le 3.5) \end{cases}$	Used to calculate hydrogen bonding energy between water and polar groups. $G(\theta) = \cos^4(\theta - 120)$ for sp2-hybridized polar atom; $G(\theta) = \cos^4(\theta - 109.5)$ for sp3-hybridized polar atom; and $G(\theta)$=0.8 for hbond between two water molecules The hydrogen bonding energy is calculated if $\theta \ge 90°$

(3) no CB atom existed within 5–7 Å to the water together with the angle defined above was <90°.

Step 7. Remove water molecules that established only weak hydrogen bonding interactions with the functional groups of the truncated protein and the ligand or produced large clashes with other atoms. Water molecules were removed if their hydrogen bonding energies were weaker than −0.9 kcal mol^{-1} or the repulsive energies were greater than 5.0 kcal mol^{-1}.

Model for water-mediated hydrogen bonding

In the above-mentioned solvated ligand rotamer approach, the added water molecules were taken as being atoms of the ligand itself, that is to say, interactions between water molecules and the ligand were not taken into account. A water-mediated hydrogen bonding energy function was developed to model the interaction between water molecules and protein atoms. Figure 2a illustrates the four geometrical parameters considered: (1) the distance *d* between the protein polar atom (donor/acceptor) and oxygen of water; (2) the angle Φ1 centered at the ligand polar atom; (3) the angle Φ2 centered at the protein polar atom; and (4) the angle Ω centered at the water oxygen atom. Relatively relaxed restrictions on angles and distance were imposed on this energy function due to the inaccuracies of the water molecule placement. The angular dependence was described by simple cutoffs, as Φ1>90°, Φ2 >90°, Ω>60°. Only the water-mediated hydrogen bonding was considered when the distance *d* was between 2.6 Å and 3.5 Å. The water-mediated hydrogen bonding energy between the water and the protein donor or acceptor atom was characterized by a CHARMM [28] type of equation, as

$$E_{hb} = E_0\left\{5\left(\frac{r_0}{r}\right)^{12} - 6\left(\frac{r_0}{r}\right)^{10}\right\} \cdot \cos^4(-109.5),$$

where *r* is the distance between the water and the protein donor/acceptor atom, r_0 is the equilibrium distance, set at 2.8 Å and E_0 is the well depth, set at 8.0 kcal mol^{-1}. Figure 2b shows that the angle Ω closest to 109.5° was used in the formula if the oxygen atom of a water molecule bound to multiple ligand polar atoms, and all water-mediated hydrogen bonding energies were calculated if the water oxygen atom made multiple hydrogen bonds with different protein donors/acceptors in the same residue. In addition to the water-mediated hydrogen bonding energy term, the water molecules interact with all protein atoms via the same linear van der Waals repulsive term used in the free energy function to avoid clashes between atoms [36].

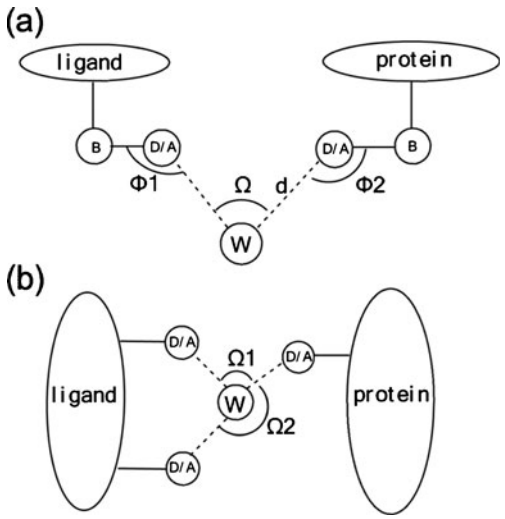


Fig. 2a,b Water-mediated hydrogen bonding model. **a** The four geometrical parameters of the model; *D/A* donor/acceptor on protein and ligand, *W* water. **b** If the oxygen atom of water bounds to multiple polar atoms of the ligand, the angle Ω most close to 109.5° is used in the formula

Results and discussion

Statistical analysis of hydrogen bonds at the protein–ligand interface

The statistical results of hydrogen bonding in the active regions of 42 protein–ligand complex structures are shown in Table 5. Only hydrogen bonds formed between t donors and acceptors from protein side-chains, ligands, and the bound waters were collected. We classified the hydrogen-bonding atoms from protein side-chains and ligands into four columns according to their buried ratios, i.e., below 25 %, from 25 % to 50 %, from 50 % to 75 %, and above 75 %, to investigate the relationship between the buried ratio of the polar atom and its ability to form hydrogen bonds, especially with the specific bound waters. The number of hydrogen bonds formed between the protein side-chains or ligands and the water molecules is not complete since many water molecules were lost in the crystal structure. However, the ratio between the number of water-mediated hydrogen bonds and the number of all hydrogen bonds formed decreased as the buried ratio of the polar atoms increased: as 90 % (75/83) for a buried ratio below 25 %, 83 % (159/191) for a buried ratio from 25 % to 50 %, 69 % (204/296) for a buried ratio from 50 % to 75 %, and 33 % (598/1795) for a buried ratio above 75 %. It should be noted that approximately 33 % of the polar atoms formed water-mediated hydrogen bonds even if their buried ratio was above 75 %, clearly indicating the great significance of bound water molecules for binding between protein and ligand. However, this phenomenon cannot be described correctly by the continuum solvent model if the explicitly bound waters at the protein–ligand interface were not taken into account. Figure 3 shows the water-mediated hydrogen bonds at protein–ligand interfaces for scaffolds 5CNA and 8XIA based on solvent accessibility. In Fig. 3a for 5CNA, the

Table 5 Statistical results of hydrogen bonds formed in active regions

Group	Buried 0–25 %		Buried 25–50 %		Buried 50–75 %		Buried 75–100 %	
	No. atom[a]	No.H-bond[b]	No. atom	No.H-bond	No. atom	No.H-bond	No. atom	No.H-bond
Lys NZ	18	7(6)	8	12(9)	14	30(20)	17	49(15)
Gln OE1	9	7(7)	6	4(1)	9	6(6)	31	40(19)
Gln NE2	3	2(2)	10	9(8)	13	11(5)	29	41(14)
Glu OE1	7	5(4)	9	8(7)	10	9(6)	42	65(28)
Glu OE2	7	4(4)	9	11(9)	9	11(8)	43	65(31)
Trp NE1	2	0(0)	3	3(3)	4	3(3)	30	26(6)
Thr OG1	3	4(4)	5	5(4)	11	18(11)	94	185(52)
Asp OD1	7	5(5)	15	28(23)	14	27(20)	89	165(64)
Asp OD2	8	6(5)	19	29(26)	13	22(19)	85	156(60)
Arg NE	9	0(0)	5	1(1)	11	8(5)	47	45(11)
Arg NH1	6	0(0)	6	6(4)	13	14(6)	47	78(18)
Arg NH2	10	8(7)	10	7(5)	11	13(6)	41	75(18)
Asn OD1	6	3(3)	10	8(7)	11	8(7)	58	90(29)
Asn ND2	5	2(2)	10	11(9)	11	12(8)	59	95(33)
Ser OG	8	9(9)	11	13(13)	12	24(18)	85	181(45)
His ND1	5	2(1)	1	1(0)	4	5(5)	45	34(8)
His NE2	2	0(0)	2	1(1)	6	4(4)	45	24(12)
Tyr OH	5	2(1)	6	7(7)	16	19(12)	54	90(38)
Ligand O^{sp2}	8	4(2)	16	23(18)	19	36(23)	84	151(58)
Ligand O^{sp3}	14	13(13)	5	3(3)	9	10(7)	63	97(28)
Ligand N^{sp2}	0	0(0)	1	1(1)	6	6(5)	38	33(8)
Ligand N^{sp3}	1	0(0)	0	0(0)	0	0(0)	2	10(3)
Total	143	83(75)	167	191(159)	226	296(204)	1128	1,795(598)

[a]Number of groups in the dataset

[b]Number of total hydrogen bonds that the group formed with all other hydrogen bonding groups, and the number in parenthesis is the number of the hydrogen bonds formed with water molecules

water molecule W45 has formed two water-mediated hydrogen bonds with OG1 atom of residue A226Thr and O2 atom of the ligand MMA238, respectively, and the buried ratios of these two polar atoms reach 50 % and 78 %. In Fig. 3b for 8XIA, the water molecule W192 has formed two water-mediated hydrogen bonds with OG1 atom of residue A90Thr and O1 atom of the ligand XLS389, respectively, and these two polar atoms are buried completely. In the context of computational protein design, the buried polar atoms are penalized because of their desolvation processes, but the water-mediated hydrogen bonding contribution is always neglected, so this will lead to the biased tendency toward the selection of the residues with less polarity at the protein–ligand interface, which is also confirmed by the sequence recapitulation test shown in the following section.

Test of solvated ligand rotamer approach

Two tests were applied to evaluate the solvated ligand rotamer approach developed in this work. One test predicted the water molecules at the protein–ligand interfaces, and compared them with those in the crystal structure. The other was used to recover the native amino acid sequences at the protein–ligand interfaces in the computational protein design framework, which tests not only the solvated ligand rotamer generation approach, but also the model to describe the water-mediated hydrogen bonding at the protein–ligand interface.

Water placement test

The solvated ligand rotamer approach was tested by using the water placement at the protein–ligand interfaces of five examples given by Huggins and Tidor [19]. The placement results are shown in Table 2. Only water molecules that lie within 3.5 Å from the ligand in the crystal structure were used to compare with waters predicted by the solvated ligand rotamer approach. The result was defined as being correct if the distance between the predicted water and that in the crystal structure was less than 2.0 Å [19]. Because all water molecules are identical and cannot be discriminated from each other, a restriction was imposed that a predicted water molecule can match with only one of the native water

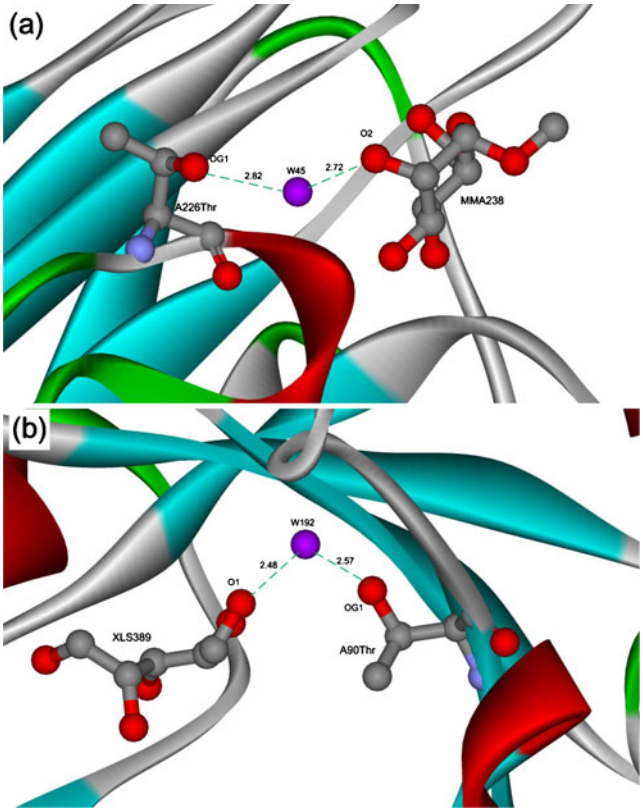


Fig. 3a,b Water-mediated hydrogen bonds at protein–ligand interface for scaffolds 5CNA and 8XIA. **a** Medium-solvent-accessible region for 5CNA. **b** Buried region for 8XIA; *solid ribbon* scaffold, *red* α-helix, *cyan* β-strand, *green* turn. The ligand and side-chain conformations of residues are shown in *ball and stick* model: *red* oxygen, *blue* nitrogen, *gray* carbon, *purple balls* water molecules (names given). *Green dotted lines* Water-mediated hydrogen bonds (numbers next to lines give distance in Å between donors and acceptors of water-mediated hydrogen bonds)

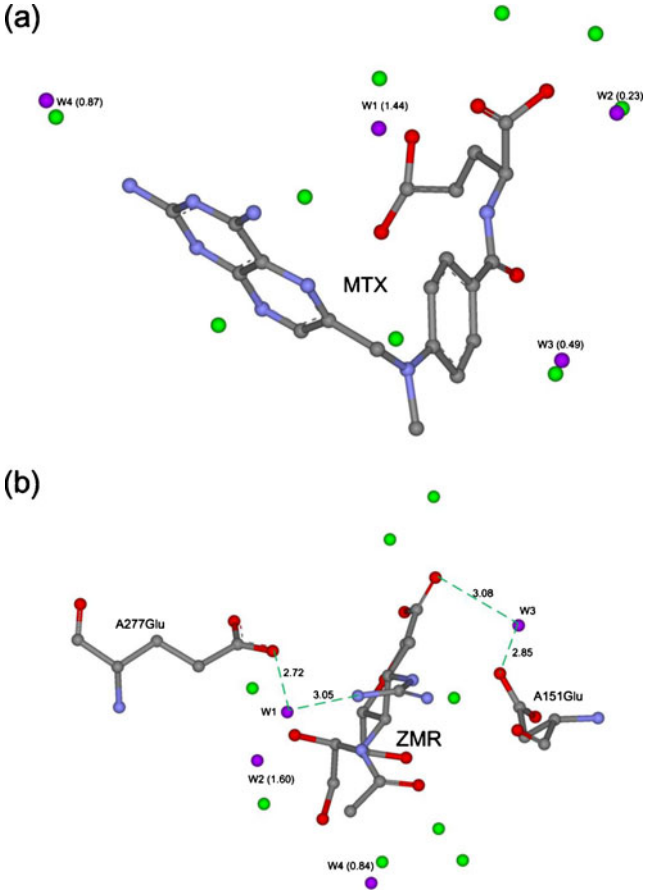


Fig. 4a,b Water placement results for 1DF7 and 1NNC. **a** 1DF7. **b** 1NNC. The ligand and the side-chain conformations of residues Glu277 and Asp151 are shown in *ball and stick models*; *red* oxygen, *blue* nitrogen, *gray* carbon. *Purple balls* Native water molecules (names given; numbers in parentheses are distances in Å between native water molecules and those predicted correctly), *green balls* predicted water molecules, *green dotted lines* water-mediated hydrogen bonds (number next to lines are distances in Å between donors and acceptors of water-mediated hydrogen bonds)

molecules. Table 2 shows that 61 % of the water molecules in the crystal structure were predicted correctly. The water placement results for 1DF7 are shown in Fig. 4a, where the four crystal water molecules around the ligand were all included in the set of the predicted water molecules, although the number of the predicted water molecules was much larger than that of the crystal waters. As shown in Table 2, there were 19 water growable positions, and most of them, i.e., ten, were removed during the solvated ligand generation process. In the case study of 1NNC, only 2 out of 4 crystal waters were predicted correctly, even though there were 22 water growable positions around the ligand, thus 8 of them were left to become the predicted waters. The reason why two crystal waters cannot be predicted correctly in this case study is shown in Fig. 4b, where the two incorrectly predicted water molecules lie very close to the side-chain functional groups of the protein but far from the ligand, therefore these water molecules were easily removed because of their collision with the pseudo spheres attached on the design sites.

Sequence recapitulation test in protein design

The five scaffolds used in the water placement tests were also applied in the sequence recapitulation test for receptor design to evaluate the whole effect of the solvated ligand rotamer approach and the water-mediated hydrogen bonding model. Three groups of calculations were implemented depending on the water models at the protein-ligand interfaces, i.e., the default implicit water model and (1) no specific water molecules concerned; (2) native water molecules; and (3) predicted water molecules. Success was recorded at the site where the wild type amino acid identity was recapitulated, and this test was based on the assumption

that the native protein had evolved to have the optimal sequence at the protein–ligand interface for binding.

Table 6 shows the amino-acid sequence recapitulation results for five examples in three different water environments at the protein–ligand interfaces. The sequence recapitulation at the protein-ligand interface was improved greatly if the specifically bound water molecules were included in the protein design model, regardless of whether the water molecules came from the crystal structures or the prediction. In fact, 44.7 % of the design sites were recovered to be the native amino acid types if the crystal or predicted water molecules were considered, but only 30.4 % were recovered without explicit waters. With respect to each case shown in Table 6, it is implied that all examples had better sequence recapitulation results with crystal water molecules than those without water molecules. Only in two examples, i.e., 1KI8 and 1VZQ, the number of the recovered design sites with predicted water molecules was less than those without water molecules, but just one less for each case. This was possibly caused by the additional water molecules around the ligand, and the correct amino acid type was eliminated because the repulsion term of the rotamers from those amino acid types was dominant due to the collision between the side-chain rotamers and the water molecules. However, in the 1NNC example, the sequence was improved greatly with the predicted water molecules even though more water molecules were placed than those in the crystal structure and only two of them were predicted correctly. For the 1DF7 and 1JIO scaffolds, the sequence recapitulation results obtained by using the predicted water model were better than those obtained by using the crystal water model. This paradox can be partly extricated by the fact that some water molecules were lost in the crystal structure. The designed amino acids at key sites for scaffold 1DF7 are shown in Fig. 5 in three different water models. Table 6 shows that the identity of residue Thr113 was predicted correctly in the model with either native or predicted water molecules, but not in the model without water. The no water result is shown Fig. 5a, where the Thr113 was mutated to His113 and the large imidazole ring of His113 occupied the position of the W4 water molecule and prevented it from forming a water-mediated hydrogen bond between the residue and the ligand. The four crystal waters shown in Fig. 4a were all predicted correctly and lie within 1.5 Å of the corresponding predicted water molecules, although the number of the latter is larger. In Fig. 5b, the conformation of Thr113 was predicted correctly either with the predicted water molecules or crystal water molecules, and the water molecule W4 formed two water-mediated hydrogen bonds with Thr113 and the ligand, which implies significance of the bound water at the protein–ligand interface. For scaffold 1JIO, the number of the predicted waters was less than that of the crystal waters, and the results are shown in Table 2 and Fig. 6. From Table 6, it can be seen

Table 6 Sequence recapitulation results

PDB	Water model type	Sequence	Correct ratio
1DF7	Without water	D27E, N28K, H30R, R32R, K53R, R60K, Y100Y, T113H	2/8
	Native water	D27E, N28K, H30Y, R32R, K53R, R60K, Y100Y, T113T	3/8
	Predicted water	D27D, N28Q, H30Y, R32R, K53R, R60Q, Y100Y, T113T	4/8
1JIO	Without water	Y75H, N89H, T92H, S93K, S171S, R185R, T291R	2/7
	Native water	Y75Y, N89S, T92N, S93K, S171S, R185R, T291R	3/7
	Predicted water	Y75Y, N89N, T92K, S93K, S171S, R185R, T291R	4/7
1KI8	Without water	H58E, K62E, E83E, Y101R, Q125Q, Y132K, R163R, Y172Y, R176K, R222R, E225R	5/11
	Native water	H58T, K62E, E83E, Y101E, Q125E, Y132K, R163R, Y172Y, R176R, R222R, E225R	5/11
	Predicted water	H58D, K62E, E83E, Y101R, Q125S, Y132K, R163R, Y172Y, R176E, R222R, E225R	4/11
1NNC	Without water	R118R, E119H, D151R, R152E, R156K, R224Q, E227D, E276R, E277H, R292K, R371R, Y406K	2/12
	Native water	R118R, E119H, D151D, R152R, R156K, R224R, E227Q, E276S, E277N, R292H, R371R, Y406K	5/12
	Predicted water	R118R, E119E, D151R, R152R, R156K, R224R, E227E, E276N, E277Q, R292S, R371E, Y406K	5/12
1VZQ[a]	Without water	H57S, Y60AY, K60FK, D189E, H192Y, D194D, S195K, S214S, Y228Y	5/9
	Native water	H57T, Y60AY, K60FK, D189E, H192R, D194D, S195D, S214S, Y228Y	5/9
	Predicted water	H57T, Y60AQ, K60FK, D189H, H192R, D194D, S195D, S214S, Y228Y	4/9

[a] Sequence indices of the second and third design sites are 60 A and 60 F given in the original PDB file

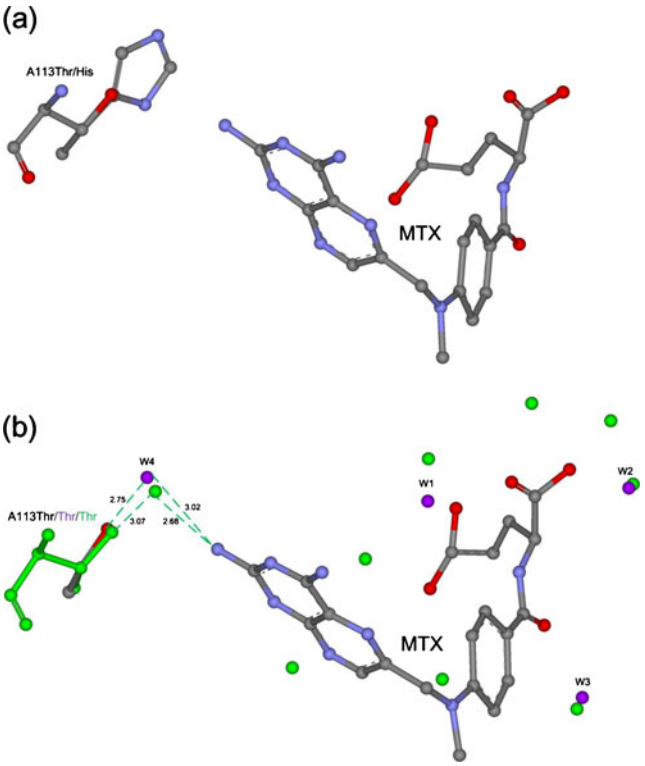


Fig. 5a,b Water placement and sequence recapitulation results for 1DF7 scaffold. **a** Sequence predicted using the model without water molecules. **b** Water placement and sequence predicted using the model with native or predicted waters, where MTX is the ligand. The native side-chain conformation of the residue at site A113 is shown in *ball and stick model*; *red* oxygen, *blue* nitrogen, *gray* carbon. *Purple balls* Native water molecules, *green balls* predicted water molecules; predicted side-chain conformations at design site A113 with native or predicted water molecules are shown in *purple* and *green*, respectively. (Note that the purple model of A113Thr is invisible as it overlaps with the green model)

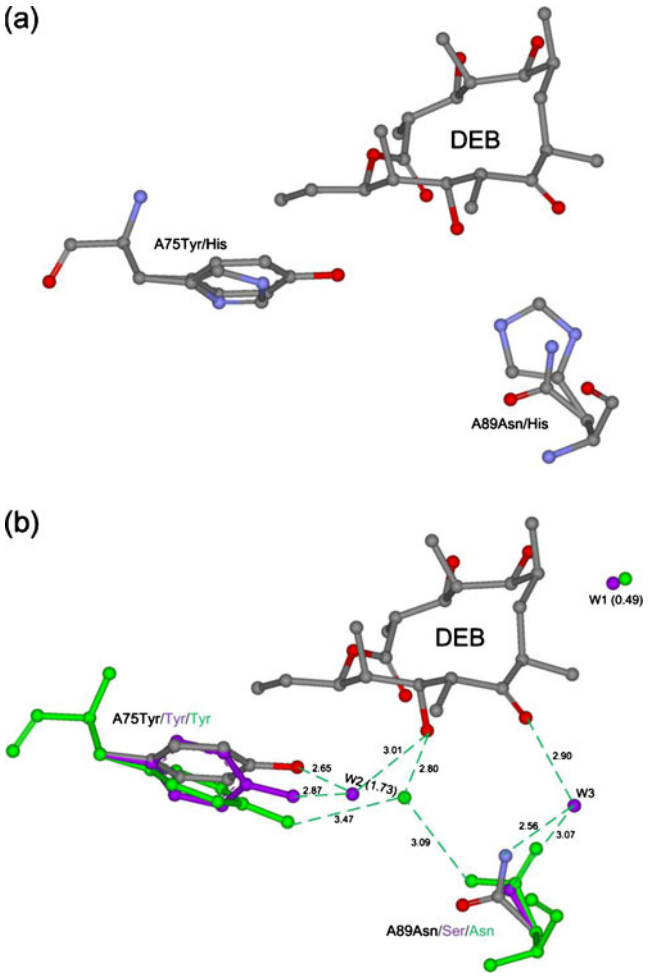


Fig. 6a,b Water placement and sequence recapitulation results for 1JIO scaffold. **a** Sequence predicted using the model without water molecule. **b** Water placement and sequence predicted using the models with native or predicted waters, where DEB is the ligand. The native side-chain conformations of residues at sites 75 and 89 are shown in *ball and stick model*; *red* oxygen, *blue* nitrogen, *gray* carbon. *Purple balls* Native water molecules (names given; numbers in parentheses are distances in Å between native water molecules and those predicted correctly), *green balls* predicted water molecules. The predicted side-chain conformations at design sites 75 and 89 with native or predicted water molecules are shown in *purple* and *green*, respectively

that the amino acid identities at design sites Tyr75 and Asn89 were both recovered with the predicted water model, but neither was recovered in the model without water. A striking result shown in Table 6 and Fig. 6b is that the amino acid at design site Asn89 was mutated to Ser89 even with the crystal water model, a possible explanation being that the repulsion caused by the inclusion of the water molecule W3, which does not exist in the predicted water model, adversely affected the selection of the amino acid type at the design site Asn89. The predicted water molecule corresponding to the crystal water W2 shown in Fig. 6b formed a water-mediated hydrogen bonding network between the Tyr75, Asn89, and the ligand, so as to stabilize the interface between the protein and the ligand. This network helped to recapitulate the native sequence at the protein-ligand interface. These results confirmed the effectiveness of the solvated ligand rotamer approach and the corresponding water-mediated hydrogen bonding model developed in this work.

Conclusions

The specific water molecules at the protein–ligand interface were analyzed statistically using a dataset of 42 protein–ligand complex structures. The results indicated that the polar side chains of the protein in the active region formed water-mediated hydrogen bonding networks even in a highly buried state. These facts confirmed the significance of bound water for binding at the protein–ligand interface, and at the same time implied the deficiency of the continuum solvent model used widely in the computational protein design field. A solvated ligand rotamer approach was developed to overcome this shortcoming as it directly

generates water molecules from the growable positions of ligand polar groups. A water-mediated hydrogen bonding model based on this approach was presented. The solvated ligand rotamer approach produces only one additional rotamer for each one in the ligand rotamer library, and thus greatly alleviates the combinatorial searching burden encountered in sequence selection. The drawback of the solvated ligand rotamer approach is that the predicted waters are strongly ligand dependent, therefore interfacial waters lying far from the ligand but still participating in the water-mediated hydrogen bonding networking at protein–ligand interface cannot be identified. The water placement results for five protein–ligand case studies showed that 61 % of the water molecules in the crystal structures were predicted correctly, although the total number of generated water molecules was larger than that of the crystal waters. Moreover, the sequence recapitulation results for these five scaffolds reflected the effect of the solved ligand rotamer approach in protein design. It is clear that the water-mediated hydrogen bonding model greatly improved the amino acid sequence prediction accuracy at the protein–ligand interface compared with the case where water molecules were not taken into account. These results not only imply the significance of the subtle influence of water-mediated hydrogen bonding in protein–ligand interface design, but also should help to overcome the disadvantages of the continuum solvent model in a tractable way.

Acknowledgments This work was supported financially by the National Natural Science Foundation of China (Grant Numbers: 20976093, 21276136), and the National High Technology Research and Development (863) Program of China (Grant Number:2012AA021204).

References

1. Poornima C, Dean P (1995) Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. J Comput Aided Mol Des 9(6):500–512
2. Poornima C, Dean P (1995) Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. J Comput Aided Mol Des 9(6):521–531
3. Poornima C, Dean P (1995) Hydration in drug design. 2. Influence of local site surface shape on water binding. J Comput Aided Mol Des 9(6):513–520
4. Ladbury JE (1996) Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. Chem Biol 3(12):973–980
5. Heine A, DeSantis G, Luz JG, Mitchell M, Wong CH, Wilson IA (2001) Observation of covalent intermediates in an enzyme mechanism at atomic resolution. Science 294(5541):369–374
6. Luque I, Freire E (2002) Structural parameterization of the binding enthalpy of small ligands. Proteins Struct Funct Bioinf 49(2):181–190
7. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11(4):371–379
8. Jaramillo A, Wodak SJ (2005) Computational protein design is a challenge for implicit solvation models. Biophys J 88(1):156–171
9. Steinbach PJ, Brooks BR (1993) Protein hydration elucidated by molecular dynamics simulation. Proc Natl Acad Sci USA 90 (19):9135
10. Levitt M, Hirshberg M, Sharon R, Laidig KE, Daggett V (1997) Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. J Phys Chem B 101(25):5051–5061
11. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O (2008) Kemp elimination catalysts by computational enzyme design. Nature 453(7192):190–195
12. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF III (2008) De novo computational design of retro-aldol enzymes. Science 319(5868):1387–1391
13. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329(5989):309–313
14. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Modeling water molecules in protein-ligand docking using GOLD. J Med Chem 48 (20):6504–6515
15. Österberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS (2002) Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins Struct Funct Bioinf 46(1):34–40
16. Rarey M, Kramer B, Lengauer T (1999) The particle concept: placing discrete water molecules during protein-ligand docking predictions. Proteins Struct Funct Bioinf 34(1):17–28
17. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Daniel T, Repasky MP, Knoll EH, Shelley M, Perry JK (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749
18. Rossato G, Ernst B, Vedani A, Smiesko M (2011) AcquaAlta: a directional approach to the solvation of ligand protein complexes. J Chem Inf Model 51(8):1867–1881
19. Huggins DJ, Tidor B (2011) Systematic placement of structural water molecules for improved scoring of protein–ligand interactions. Protein Eng Des Sel 24(10):777–789
20. Jiang L, Kuhlman B, Kortemme T, Baker D (2005) A "solvated rotamer" approach to modeling water mediated hydrogen bonds at protein–protein interfaces. Proteins Struct Funct Bioinf 58(4):893–904
21. Thanki N, Thornton J, Goodfellow J (1988) Distributions of water around amino acid residues in proteins. J Mol Biol 202(3):637–657
22. Roe SM, Teeter MM (1993) Patterns for prediction of hydration around polar residues in proteins. J Mol Biol 229(2):419–427
23. Zhu YS (2007) Mixed-integer linear programming algorithm for a computational protein design problem. Ind Eng Chem Res 46 (3):839–845
24. Luo WJ, Pei JF, Zhu YS (2010) A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity. J Mol Model 16(5):903–913
25. Lei YL, Luo WJ, Zhu YS (2011) A matching algorithm for catalytic residue site selection in computational enzyme design. Protein Sci 20(9):1566–1575
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
27. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238(5):777–793

28. Brooks BR, Bruccoleri RE, Olafson BD, Swaminathan S, Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4 (2):187–217
29. Discovery Studio (2008) Version 2.1. Accelrys: San Diego, CA
30. Baker E, Hubbard R (1984) Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 44(2):97–179
31. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M (1995) The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J Comput Chem 16(3):273–284
32. Xiang Z, Honig B (2001) Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 311(2):421–430
33. Desmet J, Maeyer MD, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356(6369):539–542
34. Goldstein RF (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 66(5):1335–1340
35. Pierce NA, Spriet JA, Desmet J, Mayo SL (2000) Conformational splitting: a more powerful criterion for dead-end elimination. J Comput Chem 21(11):999–1009
36. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97 (19):10383–10388
37. Dahiyat BI, Benjamin Gordon D, Mayo SL (1997) Automated design of the surface positions of protein helices. Protein Sci 6(6):1333–1337
38. Street AG, Mayo SL (1998) Pairwise calculation of protein solvent-accessible surface areas. Fold Des 3(4):253–258
39. Dominy BN, Brooks CL III (1999) Development of a generalized Born model parametrization for proteins and nucleic acids. J Phys Chem B 103(18):3765–3773
40. Zhang N, Zeng C, Wingreen NS (2004) Fast accurate evaluation of protein solvent exposure. Proteins Struct Funct Bioinf 57(3):565–576
41. Vizcarra CL, Zhang N, Marshall SA, Wingreen NS, Zeng C, Mayo SL (2008) An improved pairwise decomposable finite-difference Poisson–Boltzmann method for computational protein design. J Comput Chem 29(7):1153–1162
42. Creamer TP (2000) Side-chain conformational entropy in protein unfolded states. Proteins Struct Funct Bioinf 40(3):443–450
43. Lassila JK, Privett HK, Allen BD, Mayo SL (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proc Natl Acad Sci USA 103(45):16710–16715